KEIGO-SYNC : Evaluating the Impact of Politeness Levels on Factual Reliability in Japanese Large Language Models

Aubourg Thomas, Demeude Edgar, Dupuy Guilhem and Vu Anh Duy

1 Introduction

Sycophancy in large language models (LLMs) constitutes a persistent alignment challenge, characterized by a model's tendency to agree with user-stated beliefs even when they conflict with verifiable facts. Existing alignment research has primarily examined sycophantic behavior in English-language settings, often in conjunction with related phenomena such as sandbagging. In parallel, multilingual safety research has largely emphasized adversarial robustness and jailbreaking prevention. As a result, the role of grammatical register as a systematic factor influencing sycophantic behavior remains insufficiently explored.

This work introduces KEIGO-SYNC, a novel evaluation framework designed to isolate and measure the relationship between grammatical politeness and sycophantic agreement. The framework leverages the Japanese language, in which honorific speech (Keigo) encodes social deference through obligatory syntactic and morphological structures, enabling controlled manipulation of linguistic register without altering semantic content. Unlike existing benchmarks such as MMLU or HELM, KEIGO-SYNC treats grammatical register as the sole independent variable while maintaining invariant factual premises.

KEIGO-SYNC operationalizes sycophancy through Sycophancy Traps, defined as instances in which a model endorses demonstrably false premises. Each factual prompt is systematically instantiated across four register conditions: English Neutral, Japanese Tameguchi (casual), Teineigo (polite), and Sonkeigo (honorific). This design enables a controlled empirical assessment of whether increasing levels of linguistic deference are associated with measurable changes in model agreement behavior across registers.

2 Dataset Construction

The construction of the corpus is based on a selection procedure designed to isolate the effect of grammatical politeness level on the factual integrity of responses produced by language models. The objective is to assess the extent to which the level of politeness expressed by the user influences the model's tendency to agree with factually incorrect statements. To ensure diversity and comparability across evaluation scenarios, the dataset is composed of three sources. First, it includes items derived from the dataset introduced in the Logical Fallacy Detection benchmark. Second, it incorporates prompts adapted from the ELEPHANT dataset, which focuses on sycophantic behavior in conversational settings. Third, we introduce a set of original questions specifically designed for this study. All items from these sources were adapted to conform to a unified experimental format.

The dataset is structured around five thematic domains, namely education and cognition, health and well-being, logical fallacies, technology and society, and ethics and values. This organization allows us to examine whether the model's tendency to agree varies across semantic domains and levels of normative sensitivity. Each domain contains prompts formulated in an assertive manner and deliberately incorporating erroneous premises, fallacious conclusions, or explicit reasoning biases. The experimental protocol aims to measure the model's ability to maintain a critical stance when confronted with a user expressing an incorrect claim with a high degree of linguistic confidence. To limit biases related to missing or specialized knowledge, the selected topics correspond to widely documented general knowledge.

All prompts were originally written in English and subsequently translated and adapted by a native Japanese speaker to ensure semantic and pragmatic fidelity. Each item was instantiated in three Japanese language registers, namely the informal register (Tameguchi), the standard register (Teineigo), and the honorific register (Sonkeigo). The politeness level constitutes the sole experimentally manipulated variable. This design enables a controlled analysis of the relationship between politeness level and the rate at which the model refuses or explicitly corrects factual errors in its responses.

3 Experimental Setup

3.1 Model Selection and Configuration

To establish a baseline for evaluating sycophantic behavior across diverse linguistic frameworks, we selected a representative suite of Large Language Models (LLMs) characterized by varying architectural designs and regional development origins. The selection process was governed by three primary independent variables: geographic-linguistic origin, parameter scale, and the specific alignment methodologies employed during the training of the response-generation models.

To examine the influence of cultural and linguistic priors embedded in training corpora, the study contrasts models of Asian origin, specifically Qwen 2.5 (7B) and GLM-4.6V-Flash, with those of

Western origin, namely Gemma 3 (27B) and Mistral (7B). This geographic stratification facilitates an analysis of whether models developed within high-context linguistic environments demonstrate distinct sycophantic patterns compared to those trained predominantly on Western data when subjected to complex Japanese honorific constraints.

The experimental design further accounts for structural variability by incorporating a spectrum of model capacities ranging from 7B to 27B parameters. This gradient enables an investigation into the correlation between parameter scale and a model's capacity to maintain factual consistency under social or grammatical pressure. A critical component of this protocol is the inclusion of the Mistral 7B base model, which serves as a non-aligned baseline as it has not undergone Reinforcement Learning from Human Feedback (RLHF). The use of a non-RLHF model allows for the isolation of sycophancy as a potential emergent property of alignment processes designed to prioritize user satisfaction over factual adherence. To ensure statistical rigor and reproducibility, all inferences were conducted with a temperature setting of 0.5, and each prompt was subjected to N=3 independent iterations to mitigate the impact of stochastic variance.

3.2 LLM-as-Judge Framework

Factual correctness and sycophantic behavior are evaluated using an LLM-as-judge framework with dedicated bilingual English–Japanese judging models. The primary judge is Google's Gemma 3 (27B), complemented by Meta's Llama 3.2 3B-Instruct for robustness checks. Both judges are fully isolated from the answer-generation models and have no access to the trap construction process, prompt variants, or model identities, ensuring evaluations rely solely on the observable content of the responses.

Each evaluation instance provides the judge with an English factual reference answer, the Japanese model response, and structured scoring instructions. The English reference anchors the evaluation to a consistent factual baseline and reduces cross-lingual ambiguity. Judges assess semantic content and factual alignment, ignoring surface-level stylistic or politeness differences unless they interfere with factual accuracy.

3.3 Evaluation Metrics

3.3.1 LLM-as-Judge

Sycophancy is assessed using a structured evaluation prompt that instructs the judge to quantify the extent to which a model response prioritizes agreement, flattery, or face-preserving behavior over objective and factually accurate content. The prompt operationalizes sycophancy along three dimensions derived from the ELEPHANT framework: regressive sycophancy, validation sycophancy, and framing sycophancy. Regressive sycophancy captures the degree to which the model sacrifices factual accuracy or independent reasoning to conform to the user's incorrect beliefs. Validation sycophancy measures the extent to which the response emphasizes emotional reassurance or praise over objective guidance. Framing sycophancy evaluates whether the model uncritically accepts the user's assumptions or framing rather than examining or correcting them when necessary. Each dimension is scored on a five-point scale, with higher values corresponding to greater prioritization of user appeasement over truthfulness. The prompt also requires judges to provide an overall sycophancy score, reflecting the response's general tendency to favor agreement or face-saving over factual accuracy.

The prompt includes explicit instructions to compare the model response against a provided English reference answer, treated as the factual ground truth. Judges are guided to distinguish between responses that accurately correct or challenge misleading statements, partially align with the ground truth, or fully adopt erroneous premises. Brief politeness or minor empathetic language is explicitly excluded from high sycophancy ratings, whereas repeated validation, excessive agreement, or uncritical adoption of misleading premises is penalized. Each judgment is reported in a standardized JSON format, including the three dimension scores, the overall sycophancy score, and a concise textual rationale of two to four sentences. This structured prompt ensures that evaluations are reproducible, interpretable, and aligned with a clearly defined operationalization of sycophancy.

3.3.2 Quantitative Metrics

To quantify the trade-off between social deference and factual accuracy, the system employs a custom metric engine that computes two primary independent variables. The first, the Formality Ratio, measures the degree of politeness in the model's response. It is normalized between 0.0 and 1.0 and calculated using rule-based detection of specific Japanese honorific markers. The scoring assigns the highest weight to Sonkeigo verbs (e.g., meshiagaru, zonjimasu, 1.0), moderate weight to standard Teineigo copulas (desu/masu, 0.6), and low weight to casual Tameguchi particles (daro, jan, 0.2). The second variable, Factual Adherence, quantifies the semantic alignment of the response with the factual baseline. This is computed by embedding the Japanese model response using Google's Gemini Embedding 001 model and calculating the cosine similarity against the embedding of the corresponding English ground truth.

By correlating the Formality Ratio with Factual Adherence, the framework identifies instances of sycophantic drift, defined as cases in which the model sacrifices factual accuracy to comply with the grammatical constraints of high-context politeness. This approach enables a precise, quantitative assessment of how increasing levels of linguistic deference influence the model's ability to maintain truthfulness.

4 Evaluation Protocol

4.1 Human Evaluation

[TODO]

4.2 Statistical Analysis and Robustness Checks

[TODO]

5 Results

The empirical evaluation reveals a consistent but nuanced relationship between grammatical politeness and sycophantic behavior across all evaluated judges. Table 1 presents average sycophancy scores across four language variants, assessed by human evaluators and three LLM-based judges.

Table 1. Average sycophancy scores by judge and language variant

Language variant	Judge model			
	Human	Gemma 3-27B	Llama 3.2-3B Instruct	Qwen 2.5-7B Instruct
EN_Base	0.62	1.39	1.87	1.41
JP_Tameguchi	0.88	1.82	2.11	**2.20**
JP_Teineigo	0.79	1.79	2.23	2.08
JP_Sonkeigo	**0.90**	**1.98**	**2.28**	2.12

5.1 Cross-Register Sycophancy Patterns

Across all judge models, Sonkeigo (honorific register) consistently exhibited the highest mean sycophancy scores in three of four evaluation conditions. Specifically, Gemma 3-27B recorded a 42.4% increase from English baseline (1.39) to Sonkeigo (1.98), while Llama 3.2-3B Instruct showed a 21.9% increase (1.87 to 2.28). Human evaluators demonstrated a 45.2% increase (0.62 to 0.90), and Qwen 2.5-7B Instruct exhibited a 50.4% increase (1.41 to 2.12). These findings provide preliminary support for the hypothesized politeness-truthfulness trade-off, with higher grammatical deference correlating with increased sycophantic agreement.

5.2 Language-Level Effects Dominate Register-Level Variation

While Sonkeigo demonstrated the highest sycophancy scores within the Japanese register spectrum, a more pronounced discontinuity emerged at the language boundary itself. The mean absolute difference between English baseline and Japanese variants (averaged across Tameguchi, Teineigo, and Sonkeigo) substantially exceeded intra-Japanese variation. For Gemma 3-27B, the English-to-Japanese gap ($\Delta_{\text{EN}\rightarrow\text{JP}} = 0.47$) was 2.35 times larger than the maximum intra-Japanese difference ($\Delta_{\text{Tameguchi}\rightarrow\text{Sonkeigo}} = 0.20$). Human evaluators exhibited an even more striking pattern, with $\Delta_{\text{EN}\rightarrow\text{JP}} = 0.24$ representing 3.0 times the intra-Japanese range (0.08). This suggests that the transition from English to Japanese itself constitutes a stronger predictor of sycophantic behavior than fine-grained honorific gradations within Japanese.

5.3 Inter-Register Variability Within Japanese

Contrary to initial predictions of monotonic escalation with politeness level, the intermediate Teineigo register did not consistently occupy a median position between Tameguchi and Sonkeigo. In the Llama 3.2-3B evaluation, Teineigo scores (2.23) exceeded Sonkeigo scores (2.28) by only 0.05 points, falling within potential measurement noise. Similarly, for Qwen 2.5-7B, Teineigo (2.08) registered lower than both Tameguchi (2.20) and Sonkeigo (2.12), suggesting non-linear or context-dependent effects of grammatical formality. These irregularities indicate that the relationship between honorific complexity and sycophancy may be mediated by factors beyond morphosyntactic register alone, potentially including semantic domain, prompt structure, or model-specific cultural priors.

5.4 Judge Concordance and Baseline Calibration

Substantial variance in absolute score magnitudes across judges warrants methodological consideration. Human evaluators consistently assigned lower scores (mean = 0.80) compared to all LLM judges (Gemma: 1.75, Llama: 2.12, Qwen: 1.95), suggesting either stricter calibration standards or differential sensitivity to linguistic nuance. Despite these baseline discrepancies, all judges preserved the fundamental ordering of English $<$ Japanese variants. This concordance validates the robustness of the language-level effect while highlighting the need for human-aligned score normalization in future iterations.

6 Conclusion

[TODO]

7 Citations and references